Load Balancing in a Changing World: Dealing with Heterogeneity and Performance Variability

Michael Boyer, Shuai Che, Kevin Skadron
University of Virginia

Nuwan Jayasena
AMD Research

Contact: boyer@cs.virginia.edu
SRC Tasks 1607.001 and 1972.001

Abstract

Fully utilizing the power of modern heterogeneous systems requires judiciously dividing work across all of the available computational devices. Existing approaches for partitioning work generate fixed partitions even as relative device performance fluctuates due to system- or application-level changes. We present a novel dynamic approach to work partitioning that automatically responds to this performance variability in order to consistently provide good performance.

Proposed Load Balancing Algorithm

Assume we are trying to divide work across two devices:

![Device 1](image1) ![Device 2](image2)

We could statically schedule the work using a fixed partition:

But what if the optimal partition changes at runtime? We propose the following dynamic load balancing algorithm:

1. Send a fraction of the total work (a chunk) to each device:
   - ![Device 1](image3) ![Device 2](image4)
2. When a chunk completes execution:
   - ![Device 1](image5) ![Device 2](image6)
3. If not all devices have completed N chunks, launch another, larger chunk on this device:
   - ![Device 1](image7) ![Device 2](image8)
4. Otherwise, proceed to step 3.

Load Balancing Example

Sequence of data transfers and kernel executions in the application Matrix Multiplication, automatically scheduled using the algorithm described above:

![Graph](image9)

These fluctuations cause the optimal work partition to change:

![Graph](image10)

Conclusions

- Automatic load balancing can provide significant performance improvements in heterogeneous systems without requiring source code changes.
- But the optimal partition can fluctuate at runtime.
- Our proposed dynamic load balancer can effectively respond to these fluctuations, providing better average speedup (1.24x) than the best static partition (1.13x).
- The amount of performance variability determines which approach is best:
  - High variability: dynamic approach is superior
  - Low variability, static approach is superior

Future Work

- Measure efficacy of dynamic load balancing for:
  - Other system configurations
  - Other applications

Acknowledgments

This work is supported by an SRC GRC Fellowship, SRC grants 2007-HJ-1607 and 2009-HJ-1792, and equipment donated by AMD. We would also like to thank Chris Gregg and Mariabel Guevara for their helpful feedback.

Biography

Michael Boyer was awarded the Bachelor of Science Degree in Computer Engineering from Union College in 2006. As an undergraduate he received several awards, including the Loughry Prize for best senior project in Computer Engineering. He is currently a Ph.D. student at the University of Virginia, where he is researching software support for heterogeneous architectures under the guidance of Kevin Skadron. He is supported by an AMD/Mahboob Khan Fellowship from SRC.